Communications for Statistical Applications and Methods
2014, Vol. 21, No. 6, 471–486

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.6.471
Print ISSN 2287-7843 / Online ISSN 2383-4757

Forecasting Symbolic Candle Chart-Valued Time Series

Heewon Park[1,a], Fumitake Sakaori[b]

[a]Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
[b]Department of Mathematics, Faculty of Science and Engineering, Chuo University, Japan

Abstract

This study introduces a new type of symbolic data, a candle chart-valued time series. We aggregate four stock indices (*i.e.*, open, close, highest and lowest) as a one data point to summarize a huge amount of data. In other words, we consider a candle chart, which is constructed by open, close, highest and lowest stock indices, as a type of symbolic data for a long period. The proposed candle chart-valued time series effectively summarize and visualize a huge data set of stock indices to easily understand a change in stock indices. We also propose novel approaches for the candle chart-valued time series modeling based on a combination of two midpoints and two half ranges between the highest and the lowest indices, and between the open and the close indices. Furthermore, we propose three types of sum of square for estimation of the candle chart valued-time series model. The proposed methods take into account of information from not only ordinary data, but also from interval of object, and thus can effectively perform for time series modeling (*e.g.*, forecasting future stock index). To evaluate the proposed methods, we describe real data analysis consisting of the stock market indices of five major Asian countries'. We can see thorough the results that the proposed approaches outperform for forecasting future stock indices compared with classical data analysis.

Keywords: Candle chart, Symbolic data analysis, interval-valued data, time series, stock market indices of major Asian countries'.

1. Introduction

With the development of computers and data collection technology, database sizes continue to grow, and thus summaries of information and visualizations of enormous amounts of data are increasingly important. To address this issue, symbolic data analysis (Diday and Noirhomme-Fraiture, 2008) has been introduced as an extension of classical data analysis methods to take into account complete and complex information (Noirhomme-Fraiture and Brito, 2011), such as interval-valued data, histogram-valued data, multimodal data, and others. By incorporate information that cannot be represented by classical data analysis, symbolic data analysis enables effective summarization and visualization of huge databases.

Interval-valued data analysis, especially, has attracted considerable attention. Billard and Diday (2000) introduced linear regression modeling approaches to symbolic interval-valued data based on the midpoint of data intervals. To improve the model's prediction performance, Lima Neto and De Carvalho (2008) proposed a new approach based on information about the midpoint and half range of the intervals. Lima Neto *et al.* (2006) also proposed novel sum of squares methods, called NCRM1 and NCRM2, which considered the correlation between the midpoint and half range. Furthermore, Maia *et al.* (2008) introduced approaches to interval-valued time series based on the autoregressive

[1] Corresponding author: Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, Japan. E-mail: hwpark@ims.u-tokyo.ac.jp

Published 30 November 2014 / journal homepage: http://csam.or.kr

(AR) model, the autoregressive integrated moving average (ARIMA) model, the artificial neural network (ANN) model, and a hybrid ARIMA and ANN model. Arroyo *et al.* (2009) introduced various forecasting methods for a histogram time series (HTS).

We introduce a new type of symbolic data, called a candle chart-valued time series. The candle chart, which is widely used for empirically forecasting the direction of future stock indices, is composed of open, close, highest, and lowest stock indices (or prices). It implies that it is hard to understand change in stock indices by using a huge dataset, especially for the long term time series. To settle on the issue, we consider the four stock indices consisting of candle chart as a one data point, and propose a candle chart-valued time series by aggregating information of four stock indices at time t to as an one data point. We also propose approaches for fitting a time series model to symbolic candle chart-valued time series data based on midpoints and half ranges of open and close indices, and of highest and lowest indices, in line with the Centre and Range method (CRM method) for interval-valued data (Lima Neto and De Carvalho, 2008). The forecasted candle chart allows us to predict the rise or fall of the stock index.

Originalities of this study are given as,

1 Introduce a new type of symbolic data: candle chart-valued time series (CTS), which summarize complex stock indices dataset, and effectively visualize huge data.

2 Propose novel approaches to fit candle chart-valued time series model by taking account of information about not only ordinary dataset but also midpoint and half range of four stock indices.

3 Prediction of index direction is based on both statistical models and a practical method that is used in real stock market.

The proposed symbolic data, candle chart-valued time series, is a useful tool for the summarization and visualization of huge stock index dataset, since the information of the four stock indices is expressed as a one time series data, and we can clearly comprehend the market fluctuations by using the plot of candle chart-valued time series (*i.e.*, candle chart). Furthermore, the proposed methods for CTS can effectively perform for forecasting future stock indices, because the methods take into account of information from various sources, not only ordinary dataset. The candle chart-valued time series and the novel approaches for estimating the time series models are based on the ideas of interval-valued data (Lima Neto *et al.*, 2006; 2008; 2010), and we extend their approaches to the candle chart-valued time series analysis.

The rest of this article is organized as follows. We first introduce typical symbolic data and their approaches in Section 2. Section 3 presents a new type of symbolic data, candle chart-valued time series, and presents approaches for modeling of the CTS. An example using the stock indices of five major Asian countries' is presented in Section 4. Some concluding remarks are given in Section 5.

2. Symbolic Data Analysis

Suppose we have a dataset consisting of pulse rate hourly measured given in Table 1 (Billard, 2008). We can see through Table 1 that the dataset represented by classical data format may become huge, consequently, it is difficult to effectively figure out information of a patient's condition from the huge dataset.

The pulse rate dataset can be organized as daily interval (*i.e.*, [minimum, maximum]) as shown Table 2, which is typical symbolic data, called an interval-valued data. This implies that the huge

Table 1: Classical data: Pulse rate

patient	01Jul14							02Jul14							⋯
	00:00	01:00	⋯	12:00	13:00	⋯	24:00	00:00	01:00	⋯	12:00	13:00	⋯	24:00	
1	1	95	⋯	105	90	⋯	89	1	95	⋯	100	90	⋯	85	⋯
2	2	85	⋯	110	96	⋯	85	1	85	⋯	110	100	⋯	99	⋯
3	4	97	⋯	98	97	⋯	98	1	102	⋯	105	105	⋯	84	⋯
⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋮
98	90	95	⋯	103	100	⋯	87	1	83	⋯	89	90	⋯	109	⋯
99	89	95	⋯	112	110	⋯	81	1	97	⋯	99	80	⋯	116	⋯
100	85	95	⋯	101	89	⋯	100	1	101	⋯	103	95	⋯	96	⋯

Table 2: Interval-valued data: Pulse rate

patient	01Jul14	02Jul14	03Jul14	⋯
1	[85,119]	[88,119]	[84,113]	⋯
2	[80,116]	[85,120]	[83,110]	⋯
3	[81,120]	[83,115]	[88,109]	⋯
⋮	⋮	⋮	⋮	⋮
98	[83,113]	[80,121]	[81,101]	⋯
99	[89,109]	[81,120]	[87,119]	⋯
100	[80,123]	[83,116]	[85,111]	⋯

Table 3: Histogram data: Pulse rate

patient	01Jul14						
	00:00	01:00	⋯	12:00	13:00	⋯	24:00
1	1	95	⋯	105	90	⋯	89

dataset represented by classical data format can be effectively expressed by represent interval-valued data, and it leads to effective summarization and visualization of huge data.

The classical data in Table 1 can be also expressed as a formant of the histogram. Let consider the patient 1's data in 01Jul14 given as Table 3. In the viewpoint of the symbolic data analysis, the classical data can be organized as follows,

$$\mathbf{y}_i = \{p_{i1}\,[a_{i1}, b_{i1}]\,, \ldots, p_{is_i}\,[a_{is_i}, b_{is_i}]\} \tag{2.1}$$

where p_{is_i} is the relative frequency for the sub-interval $[a_{is_i}, b_{is_i}]$, $i = 1, \ldots, n$, *i.e.*, the observed histogram takes values on s_i interval for i^{th} observation (Diday and Noirhomme-Fraiture, 2008). This is histogram-valued data.

In our study, we focus on interval-valued data, and extend the approaches for interval-valued data to a novel candle chart-valued time series. We first briefly introduce the approaches for interval valued-data in the following section.

2.1. Centre and Range method

Lima Neto and De Carvalho (2008) proposed a Centre and Range method (CRM method), which is composed of information about midpoint and half range of interval on a linear regression model. The CRM method consists of midpoints $(\boldsymbol{y}^c, \boldsymbol{x}^c)$ and half ranges $(\boldsymbol{y}^r, \boldsymbol{x}^r)$ of response variable and predictor variables. The linear regression model for CRM method is based on two vectors, $\boldsymbol{w}_i = (\boldsymbol{x}_i^c, y_i^c)$ and

$\boldsymbol{r}_i = (\boldsymbol{x}_i^r, y_i^r)$ with $\boldsymbol{x}_i^c = (x_{i1}^c, \ldots, x_{ip}^c)$ and $\boldsymbol{x}_i^r = (x_{i1}^r, \ldots, x_{ip}^r)$, where

$$x_{ij}^c = \frac{x_{ij}^l + x_{ij}^u}{2}, \quad x_{ij}^r = \frac{x_{ij}^l - x_{ij}^u}{2}, \quad y_i^c = \frac{y_i^l + y_i^u}{2} \text{ and } y_i^r = \frac{y_i^l - y_i^u}{2}. \tag{2.2}$$

In the CRM method, regression models of the midpoints(y_i^c) and half range(y_i^r) are constructed as follows,

$$\begin{aligned} y_i^c &= \beta_0^c + \beta_1^c x_{i1}^c + \cdots + \beta_1^c x_{ip}^c + \varepsilon_i^c, \\ y_i^r &= \beta_0^r + \beta_1^r x_{i1}^r + \cdots + \beta_1^r x_{ip}^r + \varepsilon_i^r. \end{aligned} \tag{2.3}$$

We estimate regression coefficients $\hat{\boldsymbol{\beta}}^c = (\hat{\beta}_0^c, \hat{\beta}_1^c, \ldots, \hat{\beta}_p^c)$ by minimizing the following sum of squares,

$$S_{\text{CRM}} = \sum_{i=1}^{n} (\varepsilon_i^c)^2 + (\varepsilon_i^r)^2 . \tag{2.4}$$

2.2. NCRM1 and NCRM2 method

Lima Neto *et al.* (2006) proposed new sum of squares and linear regression modeling approaches, called as NCRM1 and NCRM2, for interval-valued data analysis. They also considered a linear regression model with $\boldsymbol{y}^c$ and $\boldsymbol{y}^r$ as response variables, and $\boldsymbol{x}_j^c$ and $\boldsymbol{x}_j^r$ for $j = 1, \ldots, p$ as predictor variables, respectively.

In NCRM1 model, Lima Neto *et al.* (2006) assumed that the regression models for midpoint and half range have same regression coefficients, and introduce the following linear regression models,

$$\begin{aligned} y_i^c &= \beta_0 + \beta_1 x_{i1}^c + \cdots + \beta_1 x_{ip}^c + \varepsilon_i^c, \\ y_i^r &= \beta_0 + \beta_1 x_{i1}^r + \cdots + \beta_1 x_{ip}^r + \varepsilon_i^r. \end{aligned} \tag{2.5}$$

The sum of square of the NCRM1 method is given by,

$$S_{\text{NCRM1}} = \sum_{i=1}^{n} (\varepsilon_i^c + \varepsilon_i^r)^2 , \tag{2.6}$$

and we estimate the regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ by minimizing the sum of squares S_{NCRM1}. The lower and upper bounds (*i.e.*, $\hat{y}^l$ and $\hat{y}^u$) are predicted as follows,

$$\hat{\boldsymbol{y}}^l = \hat{\boldsymbol{y}}^c - \hat{\boldsymbol{y}}^r \text{ and } \hat{\boldsymbol{y}}^u = \hat{\boldsymbol{y}}^c + \hat{\boldsymbol{y}}^r, \tag{2.7}$$

where $\hat{\boldsymbol{y}}^c = \boldsymbol{x}^c \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}^r = \boldsymbol{x}^r \hat{\boldsymbol{\beta}}$.

The NCRM2 method is similar to the NCRM1 method, but with different regression coefficients of y_i^c and y_i^r,

$$\begin{aligned} y_i^c &= \beta_0^c + \beta_1^c x_{i1}^c + \cdots + \beta_1^c x_{ip}^c + \varepsilon_i^c, \\ y_i^r &= \beta_0^r + \beta_1^r x_{i1}^r + \cdots + \beta_1^r x_{ip}^r + \varepsilon_i^r, \end{aligned} \tag{2.8}$$

and, the sum of square of the NCRM2 method is given as,

$$S_{\text{NCRM2}} = \sum_{i=1}^{n} (\varepsilon_i^c + \varepsilon_i^r)^2 . \tag{2.9}$$

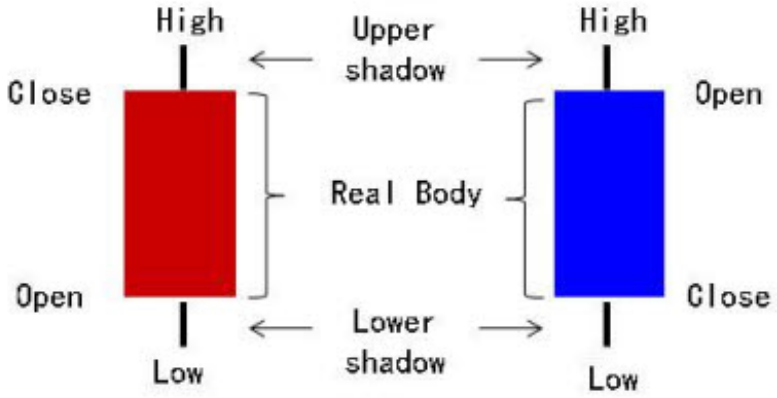

Figure 1: *Composition of the candle chart.*

We estimate the regression coefficients $\hat{\boldsymbol{\beta}}^c = (\hat{\beta}_0^c, \hat{\beta}_1^c, \ldots, \hat{\beta}_p^c)$ and $\hat{\boldsymbol{\beta}}^r = (\hat{\beta}_0^r, \hat{\beta}_1^r, \ldots, \hat{\beta}_p^r)$ by minimizing the sum of squares S_{NCRM2}.

We can see through Section 2 that the symbolic data is expressed aggregately based on information of observed data, and incorporate internal variation in the new data format. Their methods are constructed to incorporate the properties of the symbolic data. Thus, the methods outperform data analysis, since the SDA incorporate not only information of observed data but also additional information from the aggregation of the data. Furthermore, the symbolic data analysis provides effective summarization and visualization results of a huge amount of dataset.

We focus on the interval-valued data incorporating internal variation in data structure, and the approaches for interval-valued data. We consider the candle chart as a novel symbolic data, and extend the viewpoint of interval-valued data to symbolic candle chart-valued time series. The candle chart is constructed by the highest, lowest, open and close indices; however, the interval-valued data is constructed by the upper and lower bounds of data. It implies that the idea and approaches of interval-valued data can be easily extended to the CTS. We incorporate the internal variations of not only between the highest and lowest indices, but also open and close indices to time series modeling procedures, and propose approaches to estimate a time series model for CTS.

3. Candle Chart-Valued Time Series

In the following sections, we will present a candle chart-valued time series and novel approaches for fitting a time series model to CTS. In order to incorporate various information about the stock index time series, we extend the method for symbolic interval-valued time series to symbolic candle chart-valued time series via midpoint and half range of series.

Maia *et al.* (2008) introduced approaches for symbolic interval-valued time series using midpoint and half range series. In their method, two time series are considered: midpoint of interval-valued series y_t^c and half range of interval-valued series y_t^r,

$$\boldsymbol{y}_t^c = \frac{\boldsymbol{y}_{L_t} + \boldsymbol{y}_{U_t}}{2}, \quad \boldsymbol{y}_t^r = \frac{\boldsymbol{y}_{L_t} - \boldsymbol{y}_{U_t}}{2}, \quad t = 1, 2, \ldots, n, \tag{3.1}$$

where $\boldsymbol{y}_{L_t}$ is lower bound and $\boldsymbol{y}_{U_t}$ is upper bound of interval of time series $\boldsymbol{y}_t$. To forecast the interval-valued time series, Maia et al. applied the AR, ARIMA, ANN, and the hybrid models to the midpoint

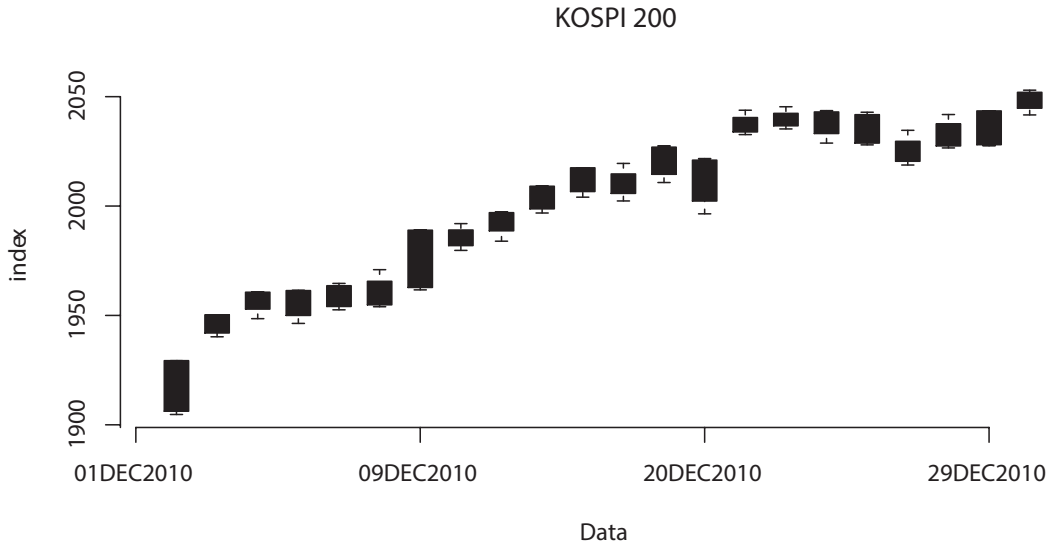

Figure 2: *Candle chart of KOSPI 200.*

Table 4: Candle chart-valued time series

	$\boldsymbol{y}_t$	$\cdots$	$\boldsymbol{y}_{t-j}$	$\cdots$	$\boldsymbol{y}_{t-p}$
$\boldsymbol{y}_1$	$\left(y^l_{1,t}, y^o_{1,t}, y^c_{1,t}, y^h_{1,t}\right)$	$\cdots$	$\left(y^l_{1,t-j}, y^o_{1,t-j}, y^c_{1,t-j}, y^h_{1,t-j}\right)$	$\cdots$	$\left(y^l_{1,t-p}, y^o_{1,t-p}, y^c_{1,t-p}, y^h_{1,t-p}\right)$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
$\boldsymbol{y}_i$	$\left(y^l_{i,t}, y^o_{i,t}, y^c_{i,t}, y^h_{i,t}\right)$	$\cdots$	$\left(y^l_{i,t-j}, y^o_{i,t-j}, y^c_{i,t-j}, y^h_{i,t-j}\right)$	$\cdots$	$\left(y^l_{i,t-p}, y^o_{i,t-p}, y^c_{i,t-p}, y^h_{i,t-p}\right)$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
$\boldsymbol{y}_n$	$\left(y^l_{n,t}, y^o_{n,t}, y^c_{n,t}, y^h_{n,t}\right)$	$\cdots$	$\left(y^l_{n,t-j}, y^o_{n,t-j}, y^c_{n,t-j}, y^h_{n,t-j}\right)$	$\cdots$	$\left(y^l_{n,t-p}, y^o_{n,t-p}, y^c_{n,t-p}, y^h_{n,t-p}\right)$

interval-valued series y^c and the half range interval-valued series y^r, respectively. They showed the superiority of the hybrid model based on the ARIMA and ANN model for time series modeling in overall.

We introduce a new type of symbolic data, candle chart-valued time series. The candle chart, which is widely used for forecasting direction of the stock index, is composed of four indexes, open ($\boldsymbol{y}^o_t$), close ($\boldsymbol{y}^c_t$), highest ($\boldsymbol{y}^h_t$) and lowest ($\boldsymbol{y}^l_t$) indexes as shown in Figure 1 (Goswamil *et al.*, 2009). The stock indices dataset may become huge, since stock price information is usually expressed by time series of four indices, open, close, highest and lowest. Thus, it is hard to understand and predict stock price (or index) fluctuation based on the huge dataset, especially for the long term time series.

To settle on the issue, we consider the candle chart as a one data point and propose a new symbolic data, candle chart-valued time series. Let $E = \{e_1, \ldots, e_n\}$ be a set of example that are describe by $p+1$ symbolic candle chart-valued time series $\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-p}$. Each example $e_i \in E (i = 1, \ldots, n)$ is represented as a candle chart-valued time series $\boldsymbol{y}_i = (\boldsymbol{y}_{i,t}, \boldsymbol{y}_{i,t-j})$ for $j = 1, 2, \ldots, p$. A variable $\boldsymbol{y}_i$ is termed a "candle chart-valued time series" composed with the open ($\boldsymbol{y}^o_i$), close ($\boldsymbol{y}^c_i$), highest ($\boldsymbol{y}^h_i$) and lowest ($\boldsymbol{y}^l_i$) stock indices, and a set E takes the values in the domain $\Re = \{(\boldsymbol{y}^o, \boldsymbol{y}^c, \boldsymbol{y}^h, \boldsymbol{y}^l), 0 < \boldsymbol{y}^l \leq \boldsymbol{y}^o, \boldsymbol{y}^c \leq \boldsymbol{y}^h\}$ as shown in Table 4.

Table 4 shows that the huge dataset of stock indices can be effectively summarized by the proposed CTS, *i.e.*, $\boldsymbol{y}_{i,t-j} = (y^l_{i,t-j}, y^o_{i,t-j}, y^c_{i,t-j}, y^h_{i,t-j})$. Furthermore, the CTS is a useful tool for the visualization

of the huge dataset of stock indices as shown in Figure 2. Figure 2 shows that the huge dataset is clearly visualized by the CTS, and thus we can efficiently predict stock indices fluctuations compared with ordinary stock indices. It may be difficult to understand the fluctuation of stock indices based on figure with ordinary four stock indices. In short, the CTS is a useful tool for forecasting stock indices, and a prime example of symbolic data, since the CTS shows the typical properties of symbolic data (*i.e.*, summarization and visualization).

3.1. Novel method for candle chart-valued time series

We proposed a novel method, called a midpoint-midpoint and range-range (MMRR) method, for fitting time series model to CTS from the viewpoint of symbolic data analysis. The MMRR method is composed of four time series: two midpoints ($\boldsymbol{y}_t^{oc_m}$, $\boldsymbol{y}_t^{hl_m}$) and two half ranges ($\boldsymbol{y}_t^{oc_r}$, $\boldsymbol{y}_t^{hl_r}$) of interval between open ($\boldsymbol{y}_t^{o}$) and close ($\boldsymbol{y}_t^{c}$) indices, and between the highest ($\boldsymbol{y}_t^{h}$) and the lowest ($\boldsymbol{y}_t^{l}$) indices, respectively,

- open-close midpoint time series: $\boldsymbol{y}_t^{oc_m}, \boldsymbol{y}_{t-1}^{oc_m}, \ldots, \boldsymbol{y}_{t-p}^{oc_m}$,
- open-close half range time series: $\boldsymbol{y}_t^{oc_r}, \boldsymbol{y}_{t-1}^{oc_r}, \ldots, \boldsymbol{y}_{t-p}^{oc_r}$,
- highest-lowest midpoint time series: $\boldsymbol{y}_t^{hl_m}, \boldsymbol{y}_{t-1}^{hl_m}, \ldots, \boldsymbol{y}_{t-p}^{hl_m}$,
- highest-lowest half range time series: $\boldsymbol{y}_t^{hl_r}, \boldsymbol{y}_{t-1}^{hl_r}, \ldots, \boldsymbol{y}_{t-p}^{hl_r}$,

where

$$\boldsymbol{y}_t^{oc_m} = \frac{\boldsymbol{y}_t^{c} + \boldsymbol{y}_t^{o}}{2}, \quad \boldsymbol{y}_t^{oc_r} = \frac{\boldsymbol{y}_t^{c} - \boldsymbol{y}_t^{o}}{2}, \quad \boldsymbol{y}_t^{hl_m} = \frac{\boldsymbol{y}_t^{h} + \boldsymbol{y}_t^{l}}{2}, \quad \boldsymbol{y}_t^{hl_r} = \frac{\boldsymbol{y}_t^{h} - \boldsymbol{y}_t^{l}}{2}, \tag{3.2}$$

where $\boldsymbol{y}_t^{oc_m} \geq 0, \boldsymbol{y}_t^{hl_m} \geq 0, \boldsymbol{y}_t^{hl_r} \geq 0, -\infty < \boldsymbol{y}_t^{oc_r} < \infty$ and $\boldsymbol{y}_t^{l} \leq \boldsymbol{y}_t^{o}, \boldsymbol{y}_t^{c} \leq \boldsymbol{y}_t^{h}$.

For the MMRR method, we apply time series models to the open-close midpoint ($\boldsymbol{y}^{oc_m}$), open-close half range ($\boldsymbol{y}^{oc_r}$), highest-lowest midpoint ($\boldsymbol{y}^{hl_m}$) and highest-lowest half range ($\boldsymbol{y}^{hl_r}$), respectively. The fitted values of these four series will be used to forecast future open, close, highest and lowest stock indices as follows,

$$\begin{aligned} \hat{\boldsymbol{y}}_t^{o} &= \hat{\boldsymbol{y}}_t^{oc_m} + \hat{\boldsymbol{y}}_t^{oc_r} \text{ and } \hat{\boldsymbol{y}}_t^{c} = \hat{\boldsymbol{y}}_t^{oc_m} - \hat{\boldsymbol{y}}_t^{oc_r}, \\ \hat{\boldsymbol{y}}_t^{h} &= \hat{\boldsymbol{y}}_t^{hl_m} + \hat{\boldsymbol{y}}_t^{hl_r} \text{ and } \hat{\boldsymbol{y}}_t^{l} = \hat{\boldsymbol{y}}_t^{hl_m} - \hat{\boldsymbol{y}}_t^{hl_r}, \end{aligned} \tag{3.3}$$

where $\hat{\boldsymbol{y}}_t^{oc_m}, \hat{\boldsymbol{y}}_t^{oc_r}, \hat{\boldsymbol{y}}_t^{hl_m}$ and $\hat{\boldsymbol{y}}_t^{hl_r}$ represent the predicted time series of open-close midpoint, open-close half range, highest-lowest midpoint and highest-lowest half range of the CTS, respectively, and we assume that $\hat{\boldsymbol{y}}_t^{oc_m} \geq 0, \hat{\boldsymbol{y}}_t^{hl_m} \geq 0, \hat{\boldsymbol{y}}_t^{hl_r} \geq 0, -\infty < \hat{\boldsymbol{y}}_t^{oc_r} < \infty$ and $\hat{\boldsymbol{y}}_t^{l} \leq \hat{\boldsymbol{y}}_t^{o}, \hat{\boldsymbol{y}}_t^{c} \leq \hat{\boldsymbol{y}}_t^{h}$. By using the MMRR method, we can take account of information about midpoint and half range of highest and lowest indices, and of open and close indices in time series modeling for stock index forecasting.

3.2. Time series model for the candle chart-valued time series

We introduce a hybrid ARIMA and ANN model which showed the outstanding performance for interval-valued time series (Hansen and Nelson, 2003) to fit the CTS. In order to explain volatility clustering of CTS, we also consider the ARIMA-ARCH model, which is widely used for financial time series having volatility clustering.

3.2.1. Hybrid model

The hybrid model, proposed by Zhang (2001), is composed of a linear component and a nonlinear component,

$$y_t = L_t + N_t, \tag{3.4}$$

where y_t is the current value of the time series at time t, L_t and N_t denote the linear and nonlinear components, respectively. The linear and nonlinear components are estimated from the data. For the hybrid model, we first apply the ARIMA (p, d, q) model for the linear component L_t,

$$\phi_p(B)(1-B)^d L_t = \theta_q(B)\varepsilon_t + \eta_t, \tag{3.5}$$

where $\phi_p(B) = (1 - \phi_1 B - \cdots \phi_p B^p)$ is the order p stationary AR operator, $\theta_q(B) = (1 - \theta_1 B - \cdots \theta_q B^q)$ is the order q invertible MA operator and B is backward shift operator as $BL_t = L_{t-1}$ and d is the order of differencing.

Then, we apply the ANN model to the residuals, which contain only the nonlinear relationship, of the ARIMA model,

$$n_t = y_t - \hat{L}_t, \tag{3.6}$$

to capture the nonlinear relation of the series using p input nodes as follows,

$$n_t = f\left(n_{t-1}, n_{t-2}, \ldots, n_{t-p}\right) + \varepsilon_t. \tag{3.7}$$

The forecasted time series $\hat{y}_t$ in the hybrid model are given by

$$\hat{y}_t = \widehat{L}_t + \widehat{N}_t, \tag{3.8}$$

where $\widehat{L}_t$ and $\widehat{N}_t$ are estimated linear and nonlinear components, respectively. For further details on this method, see Zhang (2001).

Maia *et al.* (2008) applied the hybrid model to interval-valued time series modeling, and Hansen and Nelson (2003) showed the superiority of the hybrid approach to time series modeling.

3.2.2. ARIMA-ARCH model

Financial time series often show the volatility clustering. Numerous studies of financial time series with volatility clustering have been performed using the autoregressive conditional heteroskedasticity (ARCH) model (Chen *et al.*, 2005). For time series model of CTS, we also consider the ARIMA-ARCH model to capture the volatility clustering of CTS. The ARIMA(p, d, q)-ARCH(s) model is given by

$$\begin{aligned} \phi_p(B)(1-B)^d y_t &= \theta_q(B)\varepsilon_t + \eta_t, \\ \eta_t &= \sigma_t e_t, \end{aligned} \tag{3.9}$$

where η_t are *i.i.d.* random variables with mean 0 and variance 1, which is independent of past realizations η_{t-i}, a random variable e_t is white noise process, and

$$\sigma_t = \alpha_0 + \sum_{i=1}^{s} \alpha_i \eta_{t-i}^2, \tag{3.10}$$

where $\alpha_0 > 0$, $\alpha_i > 0$ for $i = 1, \ldots, s$ and $\sum_{i=1}^{s} \alpha_1 < 1$. The differenced series $(1-B)^d y_t$ follows the general stationary ARMA (p, q) process (Wei, 2005).

3.3. Model estimation

We also propose novel approaches to estimate CTS from the viewpoint of the symbolic data analysis. In order to effectively forecasting the candle chart time series, we consider estimation methods for reflecting aggregated information of CTS in line with the method for interval-valued data (Lima Neto *et al.*, 2006).

MMRR method based on novel sum of squares

- Sum of square 1 for CTS: S_1

$$\begin{aligned} S_1 &= \sum_{i=1}^{n} \left(\varepsilon_i^{oc_m}\right)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^{oc_r}\right)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^{hl_m}\right)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^{hl_r}\right)^2 \\ &= \sum_{i=1}^{n} \left(y_{i,t}^{oc_m} - \sum_{j=1}^{p} \phi_j^{oc_m} y_{i,t-j}^{oc_m} \right)^2 + \sum_{i=1}^{n} \left(y_{i,t}^{oc_r} - \sum_{j=1}^{p} \phi_j^{oc_r} y_{i,t-j}^{oc_r} \right)^2 \\ &\quad + \sum_{i=1}^{n} \left(y_{i,t}^{hl_m} - \sum_{j=1}^{p} \phi_j^{hl_m} y_{i,t-j}^{hl_m} \right)^2 + \sum_{i=1}^{n} \left(y_{i,t}^{hl_r} - \sum_{j=1}^{p} \phi_j^{hl_r} y_{i,t-j}^{hl_r} \right)^2 . \end{aligned} \tag{3.11}$$

The S_1 is composed with respect four sum of squares errors of open-close midpoint, open-close half range, highest-lowest midpoint, and highest-lowest half range. We estimate the parameter of the time series models, $\hat{\phi}_0^{oc_m}, \ldots, \hat{\phi}_p^{oc_m}$, $\hat{\phi}_1^{oc_r}, \ldots, \hat{\phi}_p^{oc_r}$, $\hat{\phi}_1^{hl_m}, \ldots, \hat{\phi}_p^{hl_m}$ and $\hat{\phi}_1^{hl_r}, \ldots, \hat{\phi}_p^{hl_r}$ by minimize (3.11). The parameters of four variables $y_t^{oc_m}$, $y_t^{oc_r}$, $y_t^{hl_m}$, and y_t^{hlr} are independently estimated in the hybrid ARIMA and ANN model or ARIMA-ARCH model.

- Sum of square 2 for CTS: S_2

The S_2 for the MMRR method is given by

$$\begin{aligned} S_2 &= \sum_{i=1}^{n} \left(\varepsilon_i^{oc_m} + \varepsilon_i^{oc_r}\right)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^{hl_m} + \varepsilon_i^{hl_r}\right)^2 \\ &= \sum_{i=1}^{n} \left\{ \left(y_{i,t}^{oc_m} - \sum_{j=1}^{p} \phi_j^{oc_m} y_{i,t-j}^{oc_m} \right) + \left(y_{i,t}^{oc_r} - \sum_{j=1}^{p} \phi_j^{oc_r} y_{i,t-j}^{oc_r} \right) \right\}^2 \\ &\quad + \sum_{i=1}^{n} \left\{ \left(y_{i,t}^{hl_m} - \sum_{j=1}^{p} \phi_j^{hl_m} y_{i,t-j}^{hl_m} \right) + \left(y_{i,t}^{hl_r} - \sum_{j=1}^{p} \phi_j^{hl_r} y_{i,t-j}^{hl_r} \right) \right\}^2 . \end{aligned} \tag{3.12}$$

The sum of squares S_2 takes account of correlations between $y_t^{oc_m}$ and $y_t^{oc_r}$, and between $y_t^{hl_m}$ and $y_t^{hl_r}$ in estimation of time series model. In this case, the intercepts of $y_t^{oc_m}$ and $y_t^{oc_r}$ in both the hybrid and ARIMA-ARCH models, become the same because of the model identifiability. The intercept of $y_t^{hl_m}$ and $y_t^{hl_r}$ also become the same. The estimator of $\phi_0^{oc_m}, \ldots, \phi_p^{oc_m}$, $\phi_0^{oc_r}, \ldots, \phi_p^{oc_r}$, $\phi_0^{hl_m}, \ldots, \phi_p^{hl_m}$ and $\phi_0^{hl_r}, \ldots, \phi_p^{hl_r}$ can be obtained by minimize (3.12).

- Sum of square 3 for CTS: S_3

The S_3 for the MMRR method is given by

$$\begin{aligned} S_3 &= \sum_{i=1}^{n} \left(\varepsilon_i^{oc_m} + \varepsilon_i^{oc_r} + \varepsilon_i^{hl_m} + \varepsilon_i^{hl_r}\right)^2 \\ &= \sum_{i=1}^{n} \left\{ \left(y_{i,t}^{oc_m} - \sum_{j=1}^{p} \phi_j^{oc_m} y_{i,t-j}^{oc_m} \right) + \left(y_{i,t}^{oc_r} - \sum_{j=1}^{p} \phi_j^{oc_r} y_{i,t-j}^{oc_r} \right) \right. \\ &\quad \left. + \left(y_{i,t}^{hl_m} - \sum_{j=1}^{p} \phi_j^{hl_m} y_{i,t-j}^{hl_m} \right) + \left(y_{i,t}^{hl_r} - \sum_{j=1}^{p} \phi_j^{hl_r} y_{i,t-j}^{hl_r} \right) \right\}^2 . \end{aligned} \tag{3.13}$$

The sum of squares S_3 incorporates the correlation of all information of time series consisting of MMRR method, and the intercepts in all four models become the same. We can also estimate $\hat{\phi}_1^{oc_m}, \ldots, \hat{\phi}_p^{oc_m}, \hat{\phi}_1^{oc_r}, \ldots, \hat{\phi}_p^{oc_r}$, $\hat{\phi}_1^{hl_m}, \ldots, \hat{\phi}_p^{hl_m}$ and $\hat{\phi}_1^{hl_r}, \ldots, \hat{\phi}_p^{hl_r}$ by minimize (3.13).

The 4-indices method

We also introduce a 4-indices method to estimate the proposed candle chart-value time series model consisting of original open ($\boldsymbol{y}_t^o$), close ($\boldsymbol{y}_t^c$), highest ($\boldsymbol{y}_t^h$) and lowest ($\boldsymbol{y}_t^l$) indices time series. In the 4-indices method, we consider four time series models for the open ($\boldsymbol{y}_t^o$), close ($\boldsymbol{y}_t^c$), highest ($\boldsymbol{y}_t^h$) indexes, respectively. The sum of squares for the 4-indices method is given by

$$\begin{aligned} S_{4I} &= \sum_{i=1}^{n} (\varepsilon_i^o)^2 + \sum_{i=1}^{n} (\varepsilon_i^c)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^h\right)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^l\right)^2 . \\ &= \sum_{i=1}^{n} \left\{ \left(y_{i,t}^{o} - \sum_{j=1}^{p} \phi_j^{o} y_{i,t-j}^{o} \right) + \left(y_{i,t}^{c} - \sum_{j=1}^{p} \phi_j^{c} y_{i,t-j}^{c} \right) \right. \\ &\quad \left. + \left(y_{i,t}^{h} - \sum_{j=1}^{p} \phi_j^{h} y_{i,t-j}^{h} \right) + \left(y_{i,t}^{l} - \sum_{j=1}^{p} \phi_j^{l} y_{i,t-j}^{l} \right) \right\}^2 . \end{aligned} \tag{3.14}$$

The 4-indices method minimizing sum of squares S_{4I} is equivalent to fit the four independent time series models for the $\boldsymbol{y}_t^o$, $\boldsymbol{y}_t^c$, $\boldsymbol{y}_t^h$ and $\boldsymbol{y}_t^l$, respectively.

The introduced CTS and proposed methods have the following advantages:

- The candle chart valued time series enable effective summarization and visualization of huge stock indices dataset tbat consist of open, close, highest and lowest indices. Thus, we can effectively understand the stock market condition (*i.e.*, Figure 2) compared with a method based on ordinary huge dataset given as Table 4. The merit is one of the typical properties of symbolic data analysis.

- The introduced CTS incorporates various information about stock index time series, such as interval, midpoint and half range of object, not only observed object. Incorporating the various information about time series can lead to effective prediction results, and the outstanding results will be shown in next section.

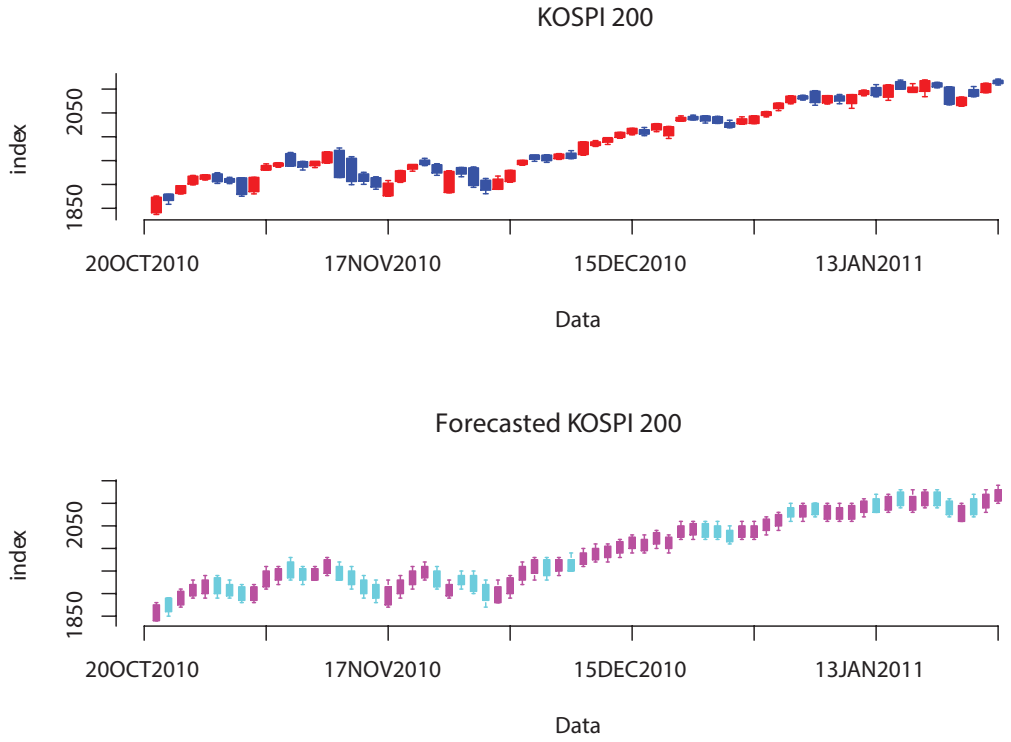

Figure 3: *Candle chart and forecasted candle chart of KOSPI 200.*

4. Applications: Stock Market Indices of Five Major Asian Countries'

We apply the proposed approaches to forecast candle chart forms in line with symbolic data analysis as shown in Figure 3, and forecast the direction of the stock index (*i.e.*, rise or fall).

In order to evaluate the proposed methods, we describe the stock market indices of five major Asian countries' (Japan, Korea, China, Singapore, and Hong Kong), which are publicly available from Korea exchange (http://eng.krx.co.kr/), based on the CTS. The databases are composed of the daily open, close, highest, and lowest indices of each of the five countries from January 2009 to April 2011. Figure 4 presents the candle chart of the stock market indices of five major Asian countries' (Korea: KOSPI 200, Japan: Nikkei 225, China: SSE, Singapore: STI, and Hong Kong: HSI).

In this study, we forecast the candle chart form via not each index but midpoints and half ranges of open and close indices, and highest and lowest indices. We estimate time series model using the dataset from January 2009 to December 2010, and then forecast the CTS from January 2011 to April 2011. In order to forecast a candle chart form, we fit the CTS by hybrid and the ARIMA-ARCH models based on the Akaike information criterion (AIC) (Akaike, 1973) as a model evaluation criterion, *i.e.*, we select the orders of ARIMA-ARCH and hybrid models that minimize AIC.

We evaluate the proposed MMRR methods based on three types of sum of squares (*i.e.*, S_1, S_2 and S_3), the method based on four indices (open, close, highest, and lowest) and using only the close index. In order to evaluate performance of the methods, we compare the root mean square error (RMSE) and the correctness of the forecasting results for the direction of the stock index (*i.e.*, a proportion of truly raising (or deceasing) daily indices among the daily stock indices estimated direction as "Up (or Down)"). In this study, we forecast the stock index direction based on the forecasting candle chart

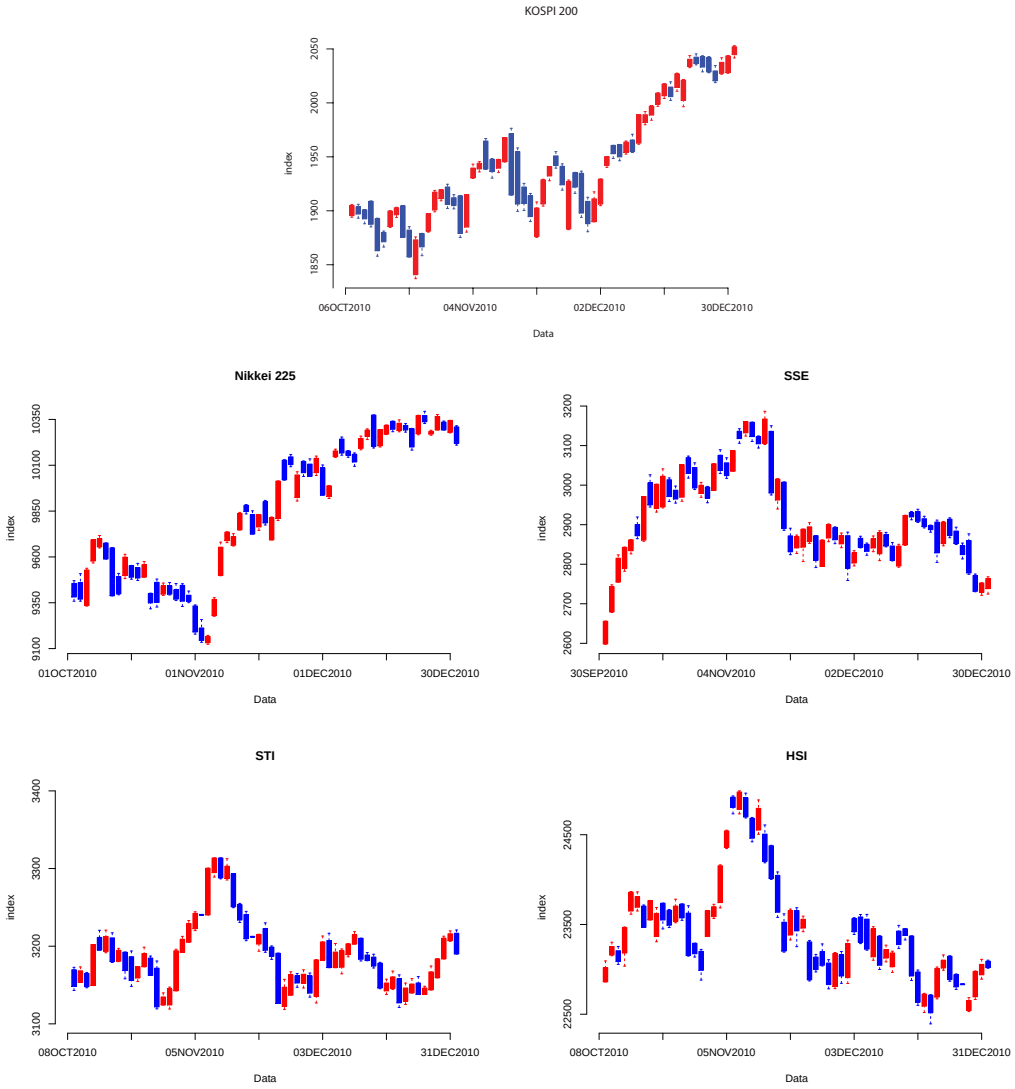

Figure 4: *Part of the candlestick chart of the stock market index of five major Asia 5 countries'.*

forms, which is widely used in real stock market, as shown in Figure 5 and below:

- The stock index will fall : 1, 3, 5, and 7
- The stock index will rise : 2, 4, 6, and 8

Table 5 shows the proportions of correctness of the forecasting results for stock index direction, where the bold number in column "Ave." (*i.e.*, average) indicates the best performance among the used methods. In the viewpoint of forecast accuracy (*i.e.*, average of proportions) and stability, the ARIMA-ARCH model with the MMRR method based on S_1 shows the superiority for forecasting the CTS as shown in column "Average" of Table 5. Furthermore, we can see through Table 5 that the

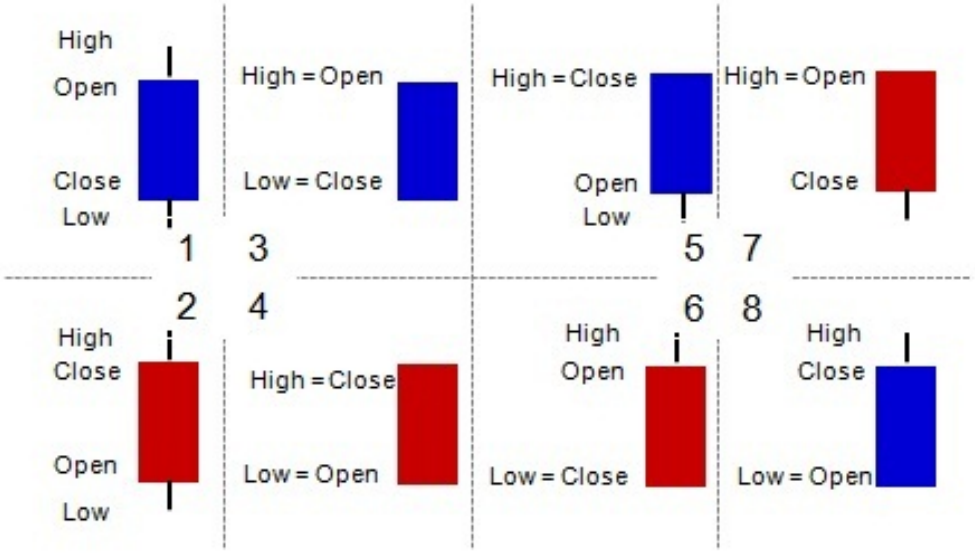

Figure 5: *Direction of stock index based on the candle chart.*

Table 5: Forecasting result of the stock index direction

			KOSPI 200		Nikkei 225		SSE		STI		HSI		**Ave.**
			Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	
ARIMA-ARCH	Close		50.1	50.0	46.4	57.9	31.6	55.6	51.7	55.6	53.6	59.5	51.2
	S_{4I}		51.9	44.4	64.5	58.1	63.5	-	57.1	48.5	58.5	50.5	55.2
	MMRR	S_1	91.2	84.8	77.4	61.1	89.2	73.1	95.6	91.9	80.0	73.5	**81.8**
		S_2	74.3	73.3	73.5	59.4	89.2	73.1	81.4	75.0	73.8	22.2	69.5
		S_3	82.1	82.8	50.0	15.4	89.2	73.1	85.4	81.8	50.0	42.0	65.2
hybrid	Close		49.0	70.4	73.3	64.7	-	100.0	100.0	100.0	81.6	72.7	79.1
	S_{4I}		96.3	80.8	80.0	64.7	68.4	39.5	100.0	100.0	81.6	75.8	78.7
	MMRR	S_1	85.7	75.0	77.4	61.1	80.0	77.8	100.0	100.0	82.4	73.5	81.3
		S_2	100.0	59.5	65.7	53.1	91.4	82.6	100.0	100.0	85.3	78.1	81.6
		S_3	100.0	42.3	36.4	20.7	91.4	82.6	66.1	100.0	85.7	78.1	70.3

methods via symbolic data analysis (*i.e.*, S_1, S_2 and S_3) show outstanding performance for predicting direction of stock indices compared with methods based on classical data analysis (*i.e.*, Close and S_{4I}). And, the sum of squares S_3 cannot perform well compared with the other proposed two sum of squares (*i.e.*, S_1 and S_2). The proposed methods based on symbolic data analysis incorporate various information about data, such as interval, midpoint and half range of objects, which cannot be represented classical data analysis, and the properties may lead to outstanding performance for indices direction prediction.

Table 6 shows the forecasting root mean square error (RMSE) of four stock indices consisting of candle chart. It can be seen through Table 6 that the sum of square S_2 shows the superior performance for predicting indices in both ARIMA-ARCH and hybrid models in overall. The S_3 shows extremely large prediction errors in some indices, and the poor prediction results lead to inefficient results for predicting direction of indices as shown in Table 5. From the Tables 5 and 6, it can be also seen that the all methods and models cannot perform well for predicting both direction and indices of the

Table 6: Root mean square error of the forecasting result

			KOSPI 200				Nikkei 225				SSE				STI				HSI			
			y^o	y^c	y^h	y^l	y^o	y^c	y^h	y^l	y^o	y^c	y^h	y^l	y^o	y^c	y^h	y^l	y^o	y^c	y^h	y^l
		S_{4I}	2.49	2.42	2.16	2.79	18.91	23.84	23.84	38.62	4.27	11.70	13.79	9.34	2.88	2.90	**3.04**	**3.33**	26.05	28.21	48.57	35.52
ARIMA		S_1	1.97	2.11	3.41	3.18	10.62	17.53	26.93	31.45	2.42	2.18	13.53	8.03	1.17	2.00	7.83	4.03	15.71	18.40	52.07	56.91
ARCH	MMRR	S_2	**0.85**	**0.82**	**1.75**	**1.34**	**6.42**	**11.54**	**26.40**	**13.72**	2.42	2.18	10.99	7.32	**0.62**	2.37	5.78	4.85	**8.07**	**10.36**	**30.65**	**17.80**
		S_3	2.15	2.07	7.55	7.39	34.79	39.92	*236.52*	*123.48*	**1.97**	**1.94**	**7.45**	**6.43**	1.36	**1.67**	3.61	3.51	14.47	14.33	53.38	28.12
		S_{4I}	1.44	0.33	5.08	4.94	0.42	0.40	1.27	*96.19*	0.18	13.31	*92.60*	*73.07*	0.31	0.31	1.15	1.17	0.35	0.33	1.20	*83.14*
hybrid		S_1	0.34	0.33	0.35	0.34	8.90	18.90	26.83	22.08	0.18	0.20	0.17	1.26	0.34	0.34	5.64	4.47	12.74	12.65	32.49	30.09
	MMRR	S_2	0.34	0.33	1.40	0.34	8.29	8.33	5.04	4.91	0.18	0.20	1.29	1.95	0.34	0.34	1.25	1.27	3.01	3.16	12.01	8.42
		S_3	0.34	0.33	1.40	0.34	29.23	34.45	*212.76*	*110.38*	0.18	0.20	2.53	6.69	0.34	0.34	0.32	1.27	7.29	7.34	37.32	22.81

Nikkei 225 and HIS compared to other countries' indices. Although the hybrid model shows outstanding performance compared with the ARIMA-ARCH model, extremely large values of RMSE are shown in not a few indices in the hybrid model (*e.g.*, Nikkei 225's $\boldsymbol{y}^h$, $\boldsymbol{y}^l$, SSE's $\boldsymbol{y}^h$, $\boldsymbol{y}^l$, and HSI's $\boldsymbol{y}^l$). Thus, we focus on the ARIMA-ARCH model. In short, the ARIMA-ARCH model with the MMRR method based on S2 outperforms for overall forecasting performances and stability as shown Table 6, where the bold numbers indicate the best performance among the four types of sum of squares in ARIMA-ARCH model.

Our methods predict the diction of the stock indices based on not only statistical strategy (*i.e.*, our method) but also the practical method that is used in real stock market (*i.e.*, candle chart form given as Figure 5). Incorporating the various information may lead to outstanding performance to predict index direction. We can expect that the proposed method will be a useful tool to predict the stock index in real market. Furthermore, our method is useful for non-specialist in statistics and financial engineering, since the proposed method based on not complex mathematics theory but practical viewpoint based on symbolic data analysis.

5. Concluding Remarks

We have introduced a new type of symbolic data, the candle chart-valued time series by aggregating open, close, highest, and lowest stock indices. We have also proposed novel approaches for fitting time series model to candle chart-valued time series, and three types of sum of squares for estimation of time series models. We can forecast the future candle chart form and direction of stock index from the fitted values of stock indices composing the candle chart. We have observed through forecasting results of the stock market indices of five major Asian countries' that the proposed approaches outperform for forecasting future stock indices compared with classical data analysis approaches.

In this study, we have focused that the candle chart can be seen as a symbolic data and introduced novel approaches to forecast index and its direction from the practical viewpoint. Thus, the theoretical parts and explanations have been omitted. Further work remains to be done towards theoretical approaches and constructing a statistical methodology for the candle chart-valued time series, such as the lasso-type regularization (Giordani, 2011) and lag weighted lasso (Park and Sakaori, 2013). As shown in Section 3, our methods cannot perform well for predicting the Nikkei 225 and HIS. In order to effectively use the proposed method, the reason for poor result should be clarified. This study can be extended to analysis of various counties' stock indices and the cause of the poor results can be investigated based on the analysis results. Furthermore, we have applied the ARMA (or ARIMA) model with the ARCH error in the study. To improve the forecast accuracy, this study can be extended

to various time series models, *i.e.*, the general autoregressive conditional heteroskedasticity (GARCH) model and the space state model.

Acknowledgement

The authors would like to thank the associate editor and anonymous reviewers for the constructive and valuable comments that improved the quality of the paper.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, B. N. Petrov, F. Csaki (*Eds.*), *Proceedings of the 2nd International Symposium on Information Theory*, Akademiai Kiado, Budapest, 267–281.

Arroyo, J., Gonzalez-Rivera, G. and Mate, C. (2009). Forecasting with interval and histogram data, Some financial applications, *Handbook of Empirical Economics and Finance*, Aman Ullah and David E. A. Giles, *eds.* Chapman and Hall/CRC 2010, 247–279.

Billard, L. (2008). Some analyses of interval data, *Journal of Computing and Information Technology*, **4**, 225–233.

Billard, L. and Diday, E. (2000). Regression Analysis for Interval-Valued Data, in *Data Analysis, Classification, and Related Methods, Studies in Classification, Data Analysis, and Knowledge Organization, eds. H. A. L. Kiers, J. P. Rassoon, P. J. F. Groenen and M. Schader*, Springer-Verlag, Berlin, 369–374.

Chen, K., Jayaprakash, C. and Yuan, B. (2005). Conditional probability as a measure of volatility clustering in financial time series, http://arxiv.org/pdf/physics/0503157v2.pdf.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley-Interscience.

Giordani, P. (2011). Linear regression analysis for interval-valued data based on the Lasso technique, *Proceeding of 58th World Statistical Congress*, Dublin, 5576–5581.

Goswamil, M. M., Bhensdadia, C. K. and Ganatra, A. P. (2009). Candlestick analysis based short term prediction of stock price fluctuation using SOM-CBR, *2009 IEEE International Advance Computing Conference*, 1448–1452.

Hansen, J. M. and Nelson, R. D. (2003). Time-series analysis with neural networks and ARIMA-neural network hybrids, *Journal of Experimental & Theoretical Artificial Intelligence* , **15**, 315–330.

Lima Neto, E. A., De Carvalho, F. A. T. and Bezerra, L. X. T. (2006). Linear regression methods to predict interval-valued Data, *Neural Networks, SBRN '06. Ninth Brazilian Symposium on*, 125–130 .

Lima Neto, E. A., De Carvalho, F. A. T. (2008). Centre and range method for fitting a linear regression model on symbolic interval data, *Computational Statistic Data Analysis*, **52**, 1500–1515.

Lima Neto, E. A. and De Carvalho, F. A. T. (2010). Constrained linear regression models for symbolic interval-valued variables, *Computational Statistic Data Analysis*, **54**, 333-347.

Maia, A. L. S., De Carvalho, F. A. T. and Ludermir, T. B. (2008). Forecasting models for interval-valued time series, *Neurocomputing*, **71**, 3344–3352.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **4**, 157–170.

Park, H. and Sakaori, F. (2013). Lag weighted lasso for time series models, *Computational Statistics*, **28**, 493–504.

Wei, W. W. S. (2005). *Time Series Analysis: Univariate and Multivariate Methods*, Addison Wesley, New York.

Zhang, G. P. (2001). Time series forecasting using a hybrid ARIMA and neural network model, *Neurocomputing*, **50**, 159–175.

Received July 6, 2014; Revised August 1, 2014; Accepted November 8, 2014